View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Edinburgh Research Archive

Automatic Transcription of Football Commentaries in the MUMIS Project

Janienke Sturm, Judith M. Kessens, Mirjam Wester, Febe de Wet, Eric Sanders, Helmer Strik

Department of Language & Speech
University of Nijmegen, The Netherlands

{Janienke.Sturm,F.deWet,E.Sanders,H.Strik}@let.kun.nl, mwester@inf.ed.ac.uk

Abstract

This paper describes experiments carried out to automatically transcribe football commentaries in Dutch, English and German for multimedia indexing. Our results show that the high levels of stadium noise in the material create a task that is extremely difficult for conventional ASR. The baseline WERs vary from 83% to 94% for the three languages investigated. Employing state-of-the-art noise robustness techniques leads to relative reductions of 9-10% WER. Application specific words such as players' names are recognized correctly in about 50% of cases. Although this result is substantially better than the overall result, it is inadequate. Much better results can be obtained if the football commentaries are recorded separately from the stadium noise. This would make the automatic transcriptions more useful for multimedia indexing.

1. Introduction

This paper describes a study into the feasibility of applying automatic speech recognition for unsupervised transcription of spoken commentaries of football matches. The study has been carried out within the framework of the EU funded IST-programme project MUMIS (MUlti-Media Indexing and Searching environment), which aimed at developing and integrating technologies to support multimedia indexing and to facilitate search and information retrieval in multimedia databases. The research domain was football matches, in particular those of the EURO-2000 championship, which took place in Belgium and the Netherlands. In this project, several sources of information about football matches are integrated, in order to be able to search the matches for certain (video frames of) events, e.g. all goals by Patrick Kluivert. One of these sources is the transcription of the spoken commentaries of the matches. Other sources of information are subtitles, tickers, and reports from newspaper, teletext, Internet, etc. The retrieval system makes use of data in different languages (Dutch, English, and German).

In earlier work [1], preliminary results on the automatic transcription of football commentaries were reported. Due to lack of training data, the experiments described in [1] were oracle recognition experiments, i.e. the lexicons and language models used were trained on the test set, instead of on an independent training set. Although these experiments provided valuable insights into what could be achieved with this type of data, the recognition results obtained were not illustrative of what can be expected in a real-life situation. Therefore, as more data became available, we carried out more realistic, non-oracle experiments. The current paper presents the results of several non-oracle experiments aimed at optimizing recognition performance on the football data.

The paper is organized as follows. Section 2 provides general information about the speech data and the speech recognizer used. In section 3, we describe the results of the experiments that were carried out. After presenting the baseline recognition results, we will show the effects of adding more data to the lexicon and the language model, distinguishing different categories of words and applying noise robust techniques. In section 4, we will discuss the results and draw some conclusions.

2. Experimental setup

2.1. Speech material

The speech material used in this study comprises recordings of television broadcasts of football matches of the EURO-2000 championship. The recordings contain spoken commentaries as well as stadium noise. The commentators' speech in all recorded matches was orthographically transcribed. Data were available in three different languages, viz. Dutch, English and German. Table 1 describes the speech data for the three languages in terms of the number of matches, the number of words and the number of utterances. An utterance in this context means a chunk of speech of about 2 to 3 seconds. The speech data were manually segmented into chunks by the transcribers.

Table 1 Statistics of the speech material

	#matches	# utts	#words
Dutch	6	14,632	39,645
English	3	7,168	35,060
German	21	41,523	122,826

2.2. Test sets

Two matches for each language were set apart as test matches to evaluate the performance of the automatic speech recognizer: *Yugoslavia vs. The Netherlands* (YugNet) and *England vs. Germany* (EngGer). Table 2 on the next page shows the number of utterances, the number of words and the total duration of the two test sets for each of the three languages.

Table 2 Statistics of the test matches

	Match	#utts	#words	duration (hh:mm:ss)
Dutch	YugNet	2,500	5,922	1:40:36
	EngGer	2,615	5,798	1:36:48
English	YugNet	2,564	10,188	1:39:34
	EngGer	3,049	13,488	1:52:04
German	YugNet	1,447	4,011	0:57:51
	EngGer	2,134	7,280	1:27:27

The test sets consist of the spoken commentaries of the two halves of the matches, including silences and some leading and trailing comments. Due to technical problems, part of the German YugNet matches could not be used, therefore the total duration of these two matches is shorter than one would expect. In the English commentaries two speakers are present, which explains the large number of words in the two English test matches.

2.3. Automatic Speech Recognition

2.3.1. Speech recognizer

The continuous speech recognition (CSR) system that was used is the Phicos system [2][3], a standard hidden Markov model (HMM) based system. Feature extraction of 8 kHz-sampled speech files is done using a 16 ms. Hamming window with a 10 ms shift. Of each speech sample 14 Mel Frequency Cepstral Coefficients (MFCCs) and their first order derivatives are calculated, which makes a total of 28 features describing the speech signal.

The acoustic models are continuous density phone level HMMs. The HMMs used in Phicos have a tripartite structure, and each of the three parts consists of two states with identical emission distributions. The transition probabilities, which allow for loops and skips, are tied over all states. 38 phone models were trained for Dutch, 40 for English and 33 for German. In addition to the sets of phone models for each language, a model for non-speech was trained. In effect the non-speech model for MUMIS data is a noise model, as all non-speech chunks contain noise. The non-speech model consists of just one state.

The phonetic transcriptions in the lexicons were mainly obtained from CELEX [4]. For those words for which no transcription was available a grapheme-to-phoneme converter was used for Dutch. For English and German, the missing phonetic transcriptions were made manually.

The language model used is a combined unigram and bigram language model. The language model was implemented as a category language model, in which prior probabilities are attached to categories of words rather than to words. Within the categories, the words have equal prior probabilities. The categories were used only for the match specific words, such as the names of the players, the referee, the coach, etc. The category LM was used in order to be able to use probabilities that apply to a whole team rather than to specific players. The reason for this is that the prior probabilities for specific players cannot be estimated on the training data.

3. Results

3.1. EXP1: Baseline

As at the start of our experiments, only very few training data were available, we carried out a baseline experiment in which for each of the three languages phone models, acoustic models and a language model were trained on one of the two test matches. The other test match was then used as an independent test set to evaluate the recognition performance. Table 3 shows the recognition performance for the two test sets in the three languages in word error rate (WER = (insertions + deletions + substitutions) / total number of words * 100%).

Table 3 Results of baseline experiment (WER)

	YugNet	EngGer
Dutch	83.3%	84.9%
English	85.7%	85.2%
German	86.8%	93.2%

Table 3 shows that the baseline error rates for all three languages are very high (83 – 94%). The WERs in Table 3 are about 20% higher than the WERs in the oracle experiments reported in [1]. This difference can be explained by the fact that in the oracle experiments both the lexicon and the language model were based on the test match, which means that the language model has a maximum fit and that there are no OOV words. In the experiments presented in this paper, the lexicon is based on a different match. Consequently, the number of OOV words that occur in the test set is much higher (see Table 4). Experiments aimed at reducing the OOV rates are described in section 3.2.

Table 4 % OOV for all test matches

	YugNet	EngGer
Dutch	21.7%	18.0%
English	12.9%	12.1%
German	23.6%	24.2%

Another explanation for the high error rates that were observed in the baseline experiment is the Signal to Noise Ratio (SNR) of the speech data. As mentioned earlier, the data used in the current experiments contain a lot of background noise, because sounds that are picked up by various microphones in the stadium are mixed with the speech recorded by the microphones of the reporters (which is standard practice for TV broadcasts of football matches). The stadium noise is far from constant: the type of sounds present varies, as does the intensity level of the sounds. The mean SNR values (and standard deviations) of the six test matches are shown in Table 5.

Table 5 Mean SNR for all test matches

	YugNet	EngGer
Dutch	9.4 (2.9)	8.2 (2.8)
English	12.1 (2.8)	10.8 (3.3)
German	19.6 (3.7)	7.1 (1.7)

Table 5 shows that the SNRs are rather low (except for the German YugNet match). The SNR is lowest for the German EngGer match, which may explain the high baseline WER that was reported for this match (Table 3). Experiments aimed at reducing the effect of noise are described in section 3.4.

3.2. EXP2: Adding data to lexicon and language model

Table 1 shows that most data was available for German (21 matches). Therefore, to decrease the OOV rate and improve the language model fit, a large amount of training data could be used for German. The language models were trained with increasing amounts of data and lexicons were based on different amounts of training data, as well. No specific selection criteria were used for the order in which extra material was added. Figure 1 shows the results for the German test sets.

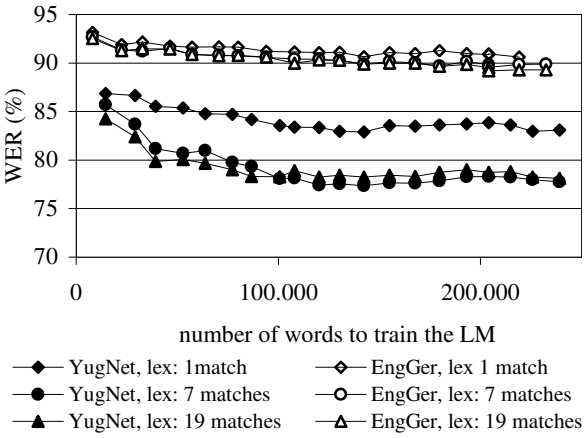

Figure 1 Results of adding data to lexicon and language model

Figure 1 shows that using more data to train the prior probabilities in the language model improves the recognition performance. Compared with the baseline performance, the relative improvements in WERs are 4-10% for YugNet and around 3% for the EngGer match. However, the improvement levels off after adding about 150.000 words.

Figure 1 also shows that using more data to generate the lexicon improves recognition performance, at least for the YugNet match. Adding more data to the lexicon increases the coverage of the lexicon. Table 6 shows the number of OOV words in the two test sets for lexicons based on 1, 7 and 19 training matches, respectively.

Table 6 %OOV in German test sets after adding extra words to the lexicon

# training matches	YugNet	EngGer
1	23.6%	24.2%
7	10.9%	10.5%
19	9.3%	6.5%

As can be seen in Table 6, the effect of adding extra data levels off; the OOV reduction going from 7 to 19 matches is smaller than the reduction going from 1 to 7 matches. This effect is also reflected in the recognition results in Figure 1: the WER improvement between 1 and 7 matches is larger than the WER improvement found between 7 and 19 matches. Furthermore, by adding extra words to the lexicon the confusability increases as well, which may hurt the recognition performance.

The improvements for the EngGer are smaller than the improvements for YugNet: adding words to the lexicon and the language model have virtually no effect. This is probably due to the level of noise in this match: the mean SNR for this match is the lowest of all matches considered.

3.3. EXP3: Word types

Usually all words are weighted equally in calculating the WERs. However, in an information retrieval task not all words are equally important. Previous research has shown that disregarding commonly occurring function words in an information retrieval task improves the retrieval performance [5]. Since function words are notoriously variable in their pronunciation [6], correct recognition of function words is also problematic. Therefore, we decided to distinguish between the following three word types: 1) function words, for instance: prepositions, pronouns, determiners etc., 2) application specific content words, and 3) match specific words - players' names. Table 7 shows the number of occurrences of each category in the six test matches.

Table 7 shows that for each of the three languages both the function words and the content words make up about 45% of the data, the names make up roughly 10% of the data.

Table 7 Number of words in each category

		Content words	Function words	Names	Total
YugNet	Dutch	2,340	2,885	695	5,922
	English	4,438	4,893	838	10,188
	German	1,966	1,541	504	4,011
EngGer	Dutch	2,214	2,852	732	5,798
	English	5,728	6,415	1,319	13,488
	German	3,457	3,097	726	7,280

We investigated how well words in the three different categories, mentioned above, were recognized. For the best testing conditions from section 3.2, separate WERs were calculated for the three categories. The results are shown in Figure 2.

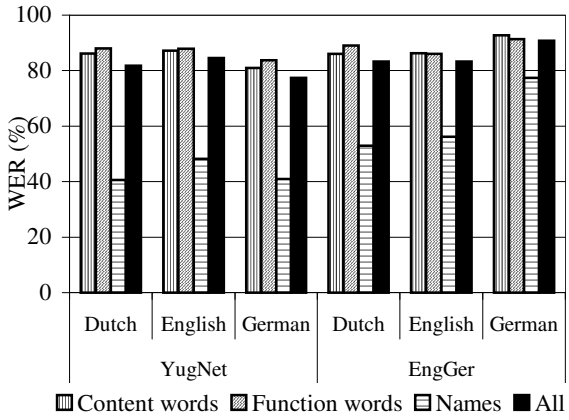

Figure 2 WERs for different categories of words (%)

Figure 2 shows that the WERs vary a lot over the three categories, ranging from 40 to 93%. In general, function words and content words are recognized about equally well. However, the WERs for the names are considerably lower.

Once again, the German EngGer match shows different results to the other tests. In this test set even the application specific category of players' names leads to a poor result.

3.4. EXP4: Noise robust techniques

It is clear that the presence of stadium noise in the speech signals (see Table 5) has a detrimental effect on the performance; therefore we applied a number of noise robust techniques in an effort to alleviate the detrimental effect of the noise. Various noise robust techniques were studied: Cepstral Mean Normalisation (CMN) [7], Mean Variance Normalisation (MVN) [8], Histogram Normalisation (HN) [9][10], and a time-domain noise reduction (TDNR) technique [11]. These techniques were tested in isolation and in various combinations; using the best testing conditions from section 3.2 (lexicon based on 7 matches, language model trained on 12 matches). Table 8 shows the best results for the German YugNet match.

Table 8 Results of noise robustness experiments

	WER
Baseline	77.7%
HN + MVN	69.9%
HN + TDNR	70.9%
HN + CMN + TDNR	70.1%

The best results were found for the following three combinations: HN + MVN, HN + TDNR, and HN + TDNR + CMN. The relative reductions in WER, compared to the WER of the baseline, for these three combinations are 10.0%, 8.9%, and 9.8%, respectively.

4. Conclusions

The aim of the research presented in this paper was to automatically transcribe spoken football commentaries. The baseline automatic speech recognition system yields WERs that vary from 84% to 94% for the different languages and test matches. The main reason that the WERs are so high is because of the extremely high level of stadium noise present in the signals that have to be transcribed, and consequently, the low SNR values of 2 to 25 dB. By applying (combinations of) noise robust techniques relative reductions of the WER of 9 -10% were achieved. Recognition performance was also improved by using more data for lexicon generation (i.e. to reduce the OOV rate) and for training the language models. However, the best WERs were found when different categories of words were distinguished: the highest WERs were found for function words, and the lowest WERs (up to 40%) for player names (which obviously are more important than function words for the given application).

Even though it has been reported in [12] that even with high WERs (around 55%) it is possible to create a usable index, the best WERs obtained in this study (even those for the player names) remain high. The lowest WERs were obtained for the parts of the matches with relatively high SNR values. Consequently, it would have been much better if 'clean' speech signals had been used, i.e. the speech signals picked up by the microphones of the reporters. In fact, initially, the speech signals are clean, and later in the broadcasting process they are mixed with the stadium noise. Automatic transcription of football commentaries, and all other broadcasts in which 'noise' is mixed with the 'clean' speech signals, could be improved substantially if the clean signals were available. Therefore, in order to improve information extraction and retrieval for broadcasts, we recommend that the 'clean' signals remain available, e.g. by storing the 'clean' signals and the stadium noise on different channels.

5. References

[1] Wester, M., Kessens, J.M. & Strik, H. (2002). "Goal-directed ASR in a multimedia indexing and searching environment (MUMIS)". *Proceedings of ICSLP-2002*, Denver, USA, pp. 1993-1996.

[2] Steinbiss, V., Ney, H., Haeb-Umbach, R., Tran, B.-H., Essen, U., Kneser, R., Oerder, M., Meier, H.-G., Aubert, X., Dugast, C. & Geller, D. (1993). "The Philips Research System for Large-Vocabulary Continuous-Speech Recognition". *Proceedings of Eurospeech '93,* Berlin, Germany, pp. 2125-2128.

[3] Strik, H., Russel, A., Heuvel, H. van den, Cucchiarini, C. & Boves, L. (1997). "Developing a spoken dialog system to automatize part of an existing inquiry system for public transport information". *Int. Journal of Speech Technology*, Vol. 2, No. 2, pp. 121-131.

[4] Baayen, R. H., Piepenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database (Release 2)* [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

[5] Ng, K. (2000). "Towards and integrated approach for spoken document retrieval". *Proceedings of ICSLP-00,* Beijing, China, pp. 672-675.

[6] Greenberg, S. (1999). "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation". *Speech Communication, 29, 159-176.*

[7] Atal, B. (1974). "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification". *J. Acoust. Soc. America*. Vol. 55, pp. 1304-1312.

[8] Viikki, A. and Laurila, K. (1998). "Cepstral domain segmental feature vector normalization for noise robust speech recognition". *Speech Communication*, Vol. 25, pp. 133-147.

[9] Molau, S., Pitz, M. & Ney, H. (2001). "Histogram normalisation in the acoustic feature space". *Proceedings of ASRU 2001*, Madonna di Campiglio, Trento, Italy.

[10] De la Torre, A., Segura, J.C., Benitez, M.C., Peinado, A.M. & Rubio, A.J. (2002). "Non-linear transformation of the feature space for robust speech recognition". *Proceedings of ICASSP 2002*, Orlando, USA.

[11] Noe, B., Sienel, J., Jouvet, D., Mauuary, L., Boves, L., Veth, J. de & Wet, F. de (2001). "Noise reduction for noise robust feature extraction for distributed speech recognition". *Proceedings of Eurospeech '01*, Aalborg, Denmark, pp. 433-436.

[12] Logan, B., Moreno, P., Van Thong, J.-M. & Whittacker, E. (2000). "An experimental study of an audio indexing system for the web". *Proceedings of ICSLP-00*, Beijing, China, pp. 676-679.